PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12P 21/06, 19/34, C12N 15/74 — A1

(11) International Publication Number: WO 00/00632

(43) International Publication Date: 6 January 2000 (06.01.00)

(21) International Application Number: PCT/US99/14776

(22) International Filing Date: 29 June 1999 (29.06.99)

(30) Priority Data:
60/090,970 29 June 1998 (29.06.98) US

(71) Applicant: PHYLOS, INC. [US/US]; 128 Spring Street, Lexington, MA 02421 (US).

(72) Inventors: WAGNER, Richard; 1007 Lowell Road, Concord, MA 01742 (US). WRIGHT, Martin, C.; 812 Memorial Drive #1105, Cambridge, MA 02139 (US). KREIDER, Brent; 4 Davis Road, Bedford, MA 01730 (US).

(74) Agent: ELBING, Karen; Clark & Elbing LLP, 176 Federal Street, Boston, MA 02110-2214 (US).

(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: METHODS FOR GENERATING HIGHLY DIVERSE LIBRARIES

(57) Abstract

Disclosed herein is a method for generating a nucleic acid library, the method involving: (a) providing a population of single-stranded nucleic acid templates, each of the templates including a coding sequence and an operably linked promoter sequence; (b) hybridizing to the population of single-stranded nucleic acid templates a mixture of substantially complementary single-stranded nucleic acid fragments, the fragments being shorter in length than the nucleic acid template; (c) contacting each of the hybridization products of step (b) with both a DNA polymerase which lacks strand displacement activity and a DNA ligase under conditions in which the fragments act as primers for the completion of a second nucleic acid strand which is substantially complementary to the nucleic acid template; and (d) contacting the products of step (c) with RNA polymerase to generate an RNA library, the library being transcribed from the second nucleic acid strand.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		

METHODS FOR GENERATING HIGHLY DIVERSE LIBRARIES

Background of the Invention

In general, this invention relates to methods for generating and altering recombinant libraries.

The ability to isolate a desired nucleic acid or amino acid sequence requires the availability of recombinant libraries of sufficient number and diversity that a particular species is represented in the library and can be identified by one or more screening techniques. Such libraries facilitate the isolation of useful compounds, including therapeutics, research diagnostics, and agricultural reagents, as well as their coding sequences.

Moreover, desirable libraries may be specifically designed to contain large numbers of possible variants of a single compound. This type of library may be used to screen for improved versions of the compound, for example, for a compound variant having optimized therapeutic efficacy.

For these or any other application, general approaches for increasing library diversity are very useful and represent an important focus of the protein design industry.

Summary of the Invention

In general, the present invention features a method for generating a nucleic acid library, the method involving: (a) providing a population of single-stranded nucleic acid templates, each of the templates including a coding sequence and an operably linked promoter sequence; (b) hybridizing to the population of single-stranded nucleic acid templates a mixture of substantially

complementary single-stranded nucleic acid fragments, the fragments being shorter in length than the nucleic acid template; (c) contacting each of the hybridization products of step (b) with both a DNA polymerase which lacks strand displacement activity and a DNA ligase under conditions in which the fragments act as primers for the completion of a second nucleic acid strand which is substantially complementary to the nucleic acid template; and (d) contacting the products of step (c) with RNA polymerase to generate an RNA library, the library being transcribed from the second nucleic acid strand.

In preferred embodiments, the method is used to introduce one or more mutations into the library; the mixture of substantially complementary single-stranded nucleic acid fragments is generated by cleaving a double-stranded nucleic acid molecule; the mixture of substantially complementary single-stranded nucleic acid fragments is generated by synthesis of random oligonucleotides; the single-stranded nucleic acid template is generated using an M13 phage carrying the nucleic acid, by digestion of one strand of a double-stranded nucleic acid template using gene VI exonuclease or lambda exonuclease, by capture of a biotinylated single nucleic acid strand using streptavidin, or by reverse transcription of RNA; the mixture of substantially complementary single-stranded nucleic acid fragments includes at least about 100 different species of nucleic acid fragments; step (b) is carried out using between 1 and approximately 1000 fragments per single-stranded nucleic acid template; a single strand of the product of step (c) is used as a nucleic acid template and steps (b) and (c) are repeated; steps (b) and (c) are repeated, using, in each round, the product of step (c) as the nucleic acid template; the method further involves providing one or more single-stranded nucleic acid fragments which form a homoduplex with the single-stranded nucleic acid template and carrying out step (b) in the presence of the homoduplex-forming fragments; the

promoter is a T7 promoter; the DNA polymerase is T4 DNA polymerase; the method further involves amplifying the product of step (c) prior to said contacting step (d); the method further involves the step of: (e) translating the RNA library to generate a protein library; the method further involves the step of: (e) linking to the 3' terminus of the coding sequence of each of substantially all of the members of the RNA library an amino acid acceptor molecule; and the method further involves the step of: (f) translating the RNA library to generate an RNA-protein fusion library.

In a second aspect, the invention features a method for reducing sequence variation in a population of nucleic acid molecules, the method involving: (a) providing a first population of single-stranded nucleic acid templates of varying sequence, each of substantially all of the templates including a coding sequence and an operably linked promoter sequence; (b) hybridizing to the members of the first population a second population of substantially complementary single-stranded nucleic acid fragments, the fragments being shorter in length than the nucleic acid template and the fragments being of substantially identical sequence; (c) contacting the hybridization products of step (b) with both a DNA polymerase which lacks strand displacement activity and a DNA ligase under conditions in which the fragments act as primers for the completion of a second nucleic acid strand which is substantially complementary to the nucleic acid template; and (d) contacting the products of step (c) with RNA polymerase to generate a population of RNA molecules, the population of RNA molecules being transcribed from the second nucleic acid strand and having reduced sequence variation relative to the first population of single-stranded nucleic acid templates.

In preferred embodiments, the method is used to remove one or more

mutations from the first population of single-stranded nucleic acid templates; step (b) involves hybridization of the first population of single-stranded nucleic acid templates to two or more different populations of substantially complementary single-stranded nucleic acid fragments; the second population of substantially complementary single-stranded nucleic acid fragments is generated by cleaving a double-stranded nucleic acid molecule; the second population of substantially complementary single-stranded nucleic acid fragments is generated by synthesis of random oligonucleotides; the single-stranded nucleic acid template is generated using an M13 phage carrying the nucleic acid, by digestion of one strand of a double-stranded nucleic acid template using gene VI exonuclease or lambda exonuclease, by capture of a biotinylated single nucleic acid strand using streptavidin, or by reverse transcription of RNA; step (b) is carried out using between 1 and approximately 1000 single-stranded nucleic acid fragments per single-stranded nucleic acid template; a single strand of the product of step (c) is used as a nucleic acid template and steps (b) and (c) are repeated; steps (b) and (c) are repeated, using, in each round, the product of step (c) as the nucleic acid template; the promoter is a T7 promoter; the DNA polymerase is T4 DNA polymerase; the method further involves amplifying the product of step (c) prior to said contacting step (d); the method further involves the step of: (e) translating the population of RNA molecules to generate a protein library; the method further involves the step of: (e) linking to the 3' terminus of the coding sequence of each of substantially all of the members of the population of RNA molecules an amino acid acceptor molecule; and the method further involves the step of: (f) translating the population of RNA molecules to generate an RNA-protein fusion library.

In a third aspect, the invention features a method for generating a

nucleic acid library, the method involving: (a) providing a population of single-stranded nucleic acid templates, each of the templates including a coding sequence; (b) providing a population of single-stranded nucleic acid molecules of varying sequence, the population of single-stranded nucleic acid templates and the population of single-stranded nucleic acid molecules of varying sequence being substantially complementary; (c) hybridizing the population of single-stranded nucleic acid templates with the population of single-stranded nucleic acid molecules of varying sequence under conditions sufficient to form duplexes; and (d) contacting the duplexes with one or more excision/repair enzymes under conditions that allow the enzymes to correct mismatched base pairs in the duplexes.

In preferred embodiments, the method further involves providing a population of single-stranded templates derived from the product of step (d) and repeating steps (c) and (d); and the steps (c) and (d) are repeated, using, in each round, a population of single-stranded templates derived from the product of step (d).

In a fourth aspect, the invention features a method for generating a nucleic acid library, the method involving: (a) providing a population of single-stranded nucleic acid templates, each of the templates including a coding sequence; (b) hybridizing to the population of single-stranded nucleic acid templates a mixture of substantially complementary single-stranded nucleic acid fragments, the fragments being shorter in length than the nucleic acid template; (c) contacting each of the hybridization products of step (b) with both a DNA polymerase which lacks strand displacement activity and a DNA ligase under conditions in which the fragments act as primers for the completion of a second nucleic acid strand which is substantially complementary to the nucleic acid template; and (d) contacting the products of step (c) with one or more

excision/repair enzymes under conditions that allow the enzymes to correct mismatched base pairs in the products.

In preferred embodiments, the method further involves providing a population of single-stranded templates derived from the product of step (d) and repeating steps (b) - (d); and steps (b) - (d) are repeated, using, in each round, a population of single-stranded templates derived from the product of step (d).

In preferred embodiments of the third and fourth aspects of the invention, the contacting with the excision/repair enzymes is carried out in vivo (for example, in a bacterial cell); the contacting with the excision/repair enzymes is carried out in vitro; the single-stranded nucleic acid template is generated using an M13 phage carrying the nucleic acid, by digestion of one strand of a double-stranded nucleic acid template using gene VI exonuclease or lambda exonuclease, by capture of a biotinylated single nucleic acid strand using streptavidin, or by reverse transcription of RNA; step (b) is carried out using between 1 and approximately 1000 single-stranded nucleic acid molecules of varying sequence or single-stranded nucleic acid fragments per single-stranded nucleic acid template; the method further involves the step of: (e) amplifying the product of step (d); each of the coding sequences is operably linked to a promoter sequence; the method further involves the step of: (e) transcribing the products of step (d) to generate an RNA library; the method further involves the step of: (f) translating the RNA library to generate a protein library; the method further involves the step of: (f) linking to the 3' terminus of the coding sequence of each of substantially all of the members of the RNA library an amino acid acceptor molecule; and the method further involves the step of: (g) translating the RNA library to generate an RNA-protein fusion library.

As used herein, by a "library" is meant at least 10^8, preferably, at least 10^{10}, more preferably, at least 10^{12}, and, most preferably, at least 10^{14} molecules having a nucleic acid and/or an amino acid component.

By a "mixture" of nucleic acid fragments is meant at least 100, preferably, at least 500, more preferably, at least 1000, and, most preferably, at least 1500 nucleic acid fragments.

By a "promoter sequence" is meant any nucleic acid sequence which provides a functional RNA polymerase binding site and which is sufficient to allow transcription of a proximal coding sequence.

By "substantially complementary" is meant that a nucleic acid strand possesses a sufficient number of nucleotides which are capable of forming matched Watson-Crick base pairs with a second nucleic acid strand to produce one or more regions of double-strandedness between the two nucleic acids. It will be understood that each nucleotide in a nucleic acid molecule need not form a matched Watson-Crick base pair with a nucleotide in an opposing strand to be substantially complementary, and that in a "mixture of substantially complementary single-stranded nucleic acid fragments," a significant fraction of the fragments will contain one or more nucleotides which form mismatches with the "single-stranded nucleic acid template."

By "strand displacement activity" is meant the ability of a polymerase or its associated helicase to disrupt base pairing between two nucleic acid strands.

By "mutation" is meant any nucleotide change and includes sequence alterations that result in phenotypic differences as well as changes which are silent.

By "duplex" is meant is meant a structure formed between two annealed nucleic acid strands in which sufficient sequence complementarity

exists between the strands to maintain a stable hybridization complex. A duplex may be either a "homoduplex," in which all of the nucleotides in the first strand appropriately base pair with all of the nucleotides in the second opposing strand, or a heteroduplex. By a "heteroduplex" is meant a structure formed between two annealed strands of nucleic acid in which one or more nucleotides in the first strand do not or cannot appropriately base pair with one or more nucleotides in the second opposing complementary strand because of one or more mismatches. Examples of different types of heteroduplexes include those which exhibit an exchange of one or several nucleotides, and insertion or deletion mutations.

By "random oligonucleotides" is meant a mixture of oligonucleotides having sequence variation at one or more nucleotide positions. Random oligonucleotides may be produced using entirely random or partially random synthetic approaches or by intentionally altering an oligonucleotide in a directed fashion.

By an "amino acid acceptor molecule" is meant any molecule capable of being added to the C-terminus of a growing protein chain by the catalytic activity of the ribosomal peptidyl transferase function. Typically, such molecules contain (i) a nucleotide or nucleotide-like moiety (for example, adenosine or an adenosine analog (di-methylation at the N-6 amino position is acceptable)), (ii) an amino acid or amino acid-like moiety (for example, any of the 20 D- or L-amino acids or any amino acid analog thereof (for example, O-methyl tyrosine or any of the analogs described by Ellman et al., Meth. Enzymol. 202:301, 1991)), and (iii) a linkage between the two (for example, an ester, amide, or ketone linkage at the 3' position or, less preferably, the 2' position); preferably, this linkage does not significantly perturb the pucker of the ring from the natural ribonucleotide conformation. Amino acid acceptors

may also possess a nucleophile, which may be, without limitation, an amino group, a hydroxyl group, or a sulfhydryl group. In addition, amino acid acceptors may be composed of nucleotide mimetics, amino acid mimetics, or mimetics of a combined nucleotide-amino acid structure.

By an amino acid acceptor being linked "to the 3' terminus" of a coding sequence is meant that the amino acid acceptor molecule is positioned after the final codon of that coding sequence. This term includes, without limitation, an amino acid acceptor molecule that is positioned precisely at the 3' end of the coding sequence as well as one which is separated from the final codon by intervening coding or non-coding sequence (for example, a sequence corresponding to a pause site). This term also includes constructs in which coding or non-coding sequences follow (that is, are 3' to) the amino acid acceptor molecule. In addition, this term encompasses, without limitation, an amino acid acceptor molecule that is covalently bonded (either directly or indirectly through intervening nucleic acid sequence) to the coding sequence, as well as one that is joined to the coding sequence by some non-covalent means, for example, through hybridization using a second nucleic acid sequence that binds at or near the 3' end of the coding sequence and that itself is bound to an amino acid acceptor molecule.

By an "RNA-protein" fusion is meant any molecule that includes a ribonucleic acid covalently bonded through an amide bond to a protein. This covalent bond is resistant to cleavage by a ribosome.

By a "protein" is meant any two or more naturally occurring or modified amino acids joined by one or more peptide bonds. "Protein," "peptide," and "polypeptide" are used interchangeably herein.

By "a population of single-stranded templates of varying sequence" is meant that the nucleic acid species of the population possess sequences

which differ at one or more nucleotide positions.

By "excision/repair enzymes" is meant any combination of enzymes sufficient to replace a mismatched base pair or loop with a standard base pair (i.e., A:T or G:C).

Brief Description of the Drawings

FIGURE 1 is a schematic representation of an exemplary fragment recombination method for generating highly diverse RNA-protein fusion libraries. In this method, the fragments are derived from a double-stranded DNA molecule into which sequence variation is introduced.

FIGURE 2 is a schematic representation of the initial steps of a second exemplary fragment recombination method. In this method, the fragments are synthetic oligonucleotides into which sequence variation is introduced.

Detailed Description

The present invention involves a number of novel and related methods for the random recombination of nucleic acid sequences, facilitating the generation of DNA, RNA, and protein libraries into which genetic alterations have been introduced. As described in more detail below, in one preferred embodiment, this technique is carried out in vitro and is used to generate traditional protein libraries or RNA-protein fusion libraries, either of which may then be used in combination with any of a variety of methods for the selection of desired proteins or peptides (or their corresponding coding sequences) from library populations. This general approach provides a means for the introduction of mutations into protein libraries in an unbiased fashion and also provides a technique by which unfavorable mutations may be removed

from a library or selected pool, or "backcrossed" out of a population of molecules during subsequent rounds of selection.

Fragment Recombination

According to one preferred method of the invention, a library is generated by the production of mutant fragments and the random recombination of these fragments with an unmutated (typically, wild-type) sequence. One example of this general approach is shown in Figure 1. As indicated in this figure, mutations are first randomly introduced into an initial double-stranded DNA sequence (termed "dsDNA(init)"). This produces a population of mutant double-stranded DNA sequences, which, in Figure 1, is termed "dsDNA(mut)." These mutations may be introduced by any technique, including PCR mutagenesis (which relies on the poor error-proofing mechanism of Taq polymerase), site-directed mutagenesis, or template-directed mutagenesis (for example, as described in Joyce and Inoue, Nucl. Acids Res. 17:171, 1989. The DNA in this mutation-containing population is subsequently fragmented using any of a variety of standard methods. For example, the DNA may be partially degraded using one or more nucleases (such as DNase I, micrococcal nuclease, restriction endonucleases, or P1 nuclease), or may be fragmented chemically using, for example, Fe•EDTA. Alternatively, mutation-containing fragments may be generated by limited nucleotide consumption during polymerization (for example, during PCR amplification), or by simple physical shearing (for example, by sonication). Preferable fragment sizes range from 25-1000 base pairs and are most preferably in the range of 50-150 base pairs.

The DNA fragments are then heated and subsequently annealed to a full-length single-stranded DNA template which is identical to the initial DNA

in sequence and which is the non-coding (or minus) strand of that DNA. In addition, in this hybridization mixture is included a second type of fragment, sometimes referred to as a "terminator fragment" (Joyce and Inoue, Nucl. Acids Res. 17:171, 1989). This terminator fragment is complementary to the 3' end of the single-stranded template and provides a polymerization primer which binds to the template in a manner that is relatively independent of the number or nature of the randomly annealed, mutation-containing fragments.

Single-stranded templates may be generated by any standard technique, for example, by using an M13 phage carrying the DNA sequence, by digestion of the coding strand of a dsDNA(init) molecule using gene VI exonuclease (Nikiforov et al, PCR Methods Appl. 3:285, 1994) or lambda exonuclease (Higuchi and Ochman, Nucl. Acids Res 17: 5865, 1989), by capture of a biotinylated DNA strand using immobilized streptavidin (Joyce and Inoue, Nucl. Acids Res. 17:171, 1989), or by reverse transcription of RNA. To carry out the template-fragment hybridization, templates are mixed with fragments using no less than one fragment molecule per template molecule and no more than approximately 1000 fragment molecules per template molecule. A low ratio of fragments to templates produces progeny strands that closely resemble the templates, whereas a higher ratio produces progeny that more closely resemble the fragments. Hybridization conditions are determined by standard methods and are designed to allow for the formation of heteroduplexes between the template and the fragments. Exemplary hybridization techniques are described, for example, in Stemmer. U.S. Patent No. 5,605,793.

Once annealed to the template, the fragments are joined together by treating with both a DNA polymerase that lacks strand displacement activity and a DNA ligase. DNA polymerases useful for this purpose include, without limitation, T4 DNA polymerase and reconstituted DNA pol II from E. coli (see,

for example, Hughes et al., J. Biol. Chem. 266:4568, 1991). Any DNA ligase (for example, T4 DNA ligase) may be utilized. In this step, the DNA duplexes may be treated first with the DNA polymerase and then with the DNA ligase, or with both enzymes simultaneously, and the step may be carried out, for example, as described in Joyce and Inoue (Nucl. Acids Res. 17:711, 1989). As shown in Figure 1, this step generates a population of double-stranded DNAs (termed "dsDNA(lib)), each member of which includes one strand typically having one or more introduced mutations. Because both the mutations initially introduced and the number and nature of the fragments annealed are random, different duplexes in the population contain different mutant sequences.

An alternative to this general approach for generating a double-stranded DNA library is shown in Figure 2. By this alternative approach, single-stranded oligonucleotide fragments are synthesized which correspond to portions of the coding strand of an initial double-stranded DNA molecule. These oligonucleotide fragments preferably range from 5-2000 nucleotides, and most preferably range from 20-100 nucleotides in length and are generated, for example, using any standard technique of nucleic acid synthesis. These oligonucleotides may be synthesized with completely random or semi-random mutations by any standard technique. Preferably, such oligonucleotides include up to 3 introduced mismatches per 20 nucleotide segment and are devoid of in frame stop codons. In addition, in certain cases, it may be desirable or necessary to increase the hybridization potential of the oligonucleotide through the introduction of non-natural, affinity-enhancing base pairs, such as C-5 propyne uridine or C-5 propyne cytidine. These techniques are described, for example, in Wagner et al., Science 260:1510, 1993.

These mutation-containing oligonucleotide fragments are next annealed to single-stranded templates which, as above, are full-length strands

identical in sequence to the non-coding (or minus) strand of the initial DNA. The fragments are joined together using DNA polymerase and DNA ligase, also as described above, to create a double-stranded DNA library (dsDNA(lib)). Again, this library contains a population of duplex molecules, containing an array of different coding strands having mutations which differ in number, position, and identity.

If desired, the above steps may be repeated, for either the fragment or the oligonucleotide approach, to introduce varying numbers of mutations into a DNA molecule. In particular, the mutated strands become the initial single-stranded templates, and mutant fragments or oligonucleotides are annealed to those strands and polymerized and ligated.

Methods for Backcrossing

As generally described above, the methods of the invention are used to introduce mutations into an initial DNA sequence. In addition, these techniques may be used to remove or reduce in frequency undesirable mutations from a DNA library. According to this approach, following fragmentation of the dsDNA(mut), oligonucleotides of wild-type sequence or specific fragments of unmutated or wild-type DNA (wtDNA) may be added to the single-stranded template together with the dsDNA(mut) fragments or oligonucleotides. The fragments are strand-separated (if necessary), annealed to the full-length single-stranded template, and joined together using DNA polymerase and DNA ligase, as described above. The use of a high concentration of unmutated oligonucleotide or fragment, relative to the corresponding mutant fragment, allows for the generation of libraries in which undesirable mutations are minimized or eliminated.

In addition, this approach may be used with existing mutation-

containing libraries to similarly decrease or eliminate undesirable sequences. This approach involves an initial library having mutant sequences and the annealing, polymerization, and ligation of fragments or oligonucleotides of wild-type sequence, as generally described above.

RNA, Protein, and RNA-Protein Libraries

In one preferred embodiment of the invention, the DNA libraries described above further include an RNA polymerase binding site, for example, for T7 or SP6 polymerase. Such binding sites are described, for example, in Milligan et al., Proc. Natl. Acad. Sci. USA 87:696, 1990. This site is positioned upstream of the coding sequence at a location which allows for transcription of the sequence. Typically such sites are located at between 5-2000 base pairs upstream of the coding sequence.

Libraries containing RNA polymerase binding sites may be altered as described above. Following polymerization and ligation, the dsDNA(lib) may be transcribed directly, for example, using an in vitro transcription system, to generate an RNA library. Alternatively, the dsDNA(lib) may be transcribed and translated directly, for example, using in vitro transcription and translation systems, to generate a protein library. Exemplary in vitro transcription systems and in vitro translation systems include T7 transcription systems, and rabbit reticulocyte, wheat germ, yeast, and E. coli translation systems.

If desired, the number of copies of each RNA or protein in the library may be increased by including a strand-specific amplification step prior to transcription. For example, PCR amplification may be carried out by incorporating unique primer-binding sequences are incorporated into the mutant strand during the polymerization and ligation steps. These sequences may be incorporated as either mismatches or sequence extensions at one or

both ends of the DNA, allowing amplification of the newly-synthesized strand without amplification of the template strand. Alternatively, linear amplification can be achieved by multiple cycles of annealing and extension of a single oligonucleotide primer that is complementary to the 3' end of the newly-synthesized strand. Subsequent PCR and transcription steps produce a majority of RNA corresponding to mutant sequences with only a small proportion of template-derived sequences.

In one preferred approach, the above methods for introducing mutations or for backcrossing out undesirable mutations may be used to produce highly diverse RNA-protein libraries. Such libraries may be constructed by ligating linkers containing a non-hydrolyzable amino acid acceptor molecule, such as puromycin, to the 3' termini of the RNAs in a library (for example, produced as described above). Exemplary techniques for generating RNA-protein fusions are described, for example, in Szostak et al., U.S.S.N. 09/007,005; and Roberts et al., Proc. Natl. Acad. Sci. USA 94:12297, 1997. Subsequent translation of these RNAs generates a library of RNA-protein fusion molecules that may subsequently be used in in vitro selection experiments.

In addition, if desired, RNA or RNA-protein fusion molecules, once selected, may be used as templates in standard PCR reactions to obtain the corresponding coding sequence. Thus, this method provides a means for carrying out fragment recombination, molecular backcrossing, selection of proteins and/or peptides, and selection of their corresponding coding sequences, all in an in vitro system.

Excision/Repair

In addition to fragment recombination approaches, excision/repair

may also be used to alter library sequences. This approach may be used to generate DNA, RNA, and RNA-protein fusion libraries. This technique relies on the fact that the dsDNA(lib)s, produced by any of the methods described above, by their nature, contain a certain number of mismatched base pairs. To generate diversity in the library sequences, these mismatches are repaired in vitro by excision/repair enzymes. This may be carried out using any excision repair system (for example, as described in Jaiswal et al., Nucl. Acids Res. 26:2184, 1998; or Fortini et al., Biochemistry 37:3575, 1998.

Alternatively, the excision/repair step may be carried out by transforming a dsDNA(lib) into a bacterial or yeast strain and exploiting the bacterial or yeast repair systems in vivo. Again, this step may be carried out by transforming the library into any standard in vivo excision/repair system. Exemplary systems are described, without limitation, in Campbell et al., Mutat. Res. 211:181, 1989; Bishop and Kolodner, Mol. Cell Biol. 6:3401, 1986; Fishel et al., J. Mol. Biol. 188:147, 1986; and Westmoreland et al., Genetics 145:29, 1997.

Because the above repair processes are random, this excision/repair method sometimes results in the introduction of mutations into a library sequence and at other times results in the backcrossing of wild-type sequence alterations into the coding strand.

In an alternative to the above approaches, in vitro or in vivo excision/repair may also be used directly to generate diverse libraries using as a substrate a mixture of dsDNA(mut) (for example, produced as described above) and dsDNA(init) or wtDNA. In this technique, the mixture is strand-separated and reannealed, and is then either incubated in vitro with excision/repair enzymes or transformed into bacteria to utilize the bacterial excision/repair system (for example, as described above). In this manner, mutations may be

randomly introduced into a sequence, and wild-type sequences may be backcrossed into dsDNA(mut) molecules.

What is claimed is:

Claims

1. A method for generating a nucleic acid library, said method comprising:

(a) providing a population of single-stranded nucleic acid templates, each of said templates comprising a coding sequence and an operably linked promoter sequence;

(b) hybridizing to said population of single-stranded nucleic acid templates a mixture of substantially complementary single-stranded nucleic acid fragments, said fragments being shorter in length than said nucleic acid template;

(c) contacting each of the hybridization products of step (b) with both a DNA polymerase which lacks strand displacement activity and a DNA ligase under conditions in which said fragments act as primers for the completion of a second nucleic acid strand which is substantially complementary to said nucleic acid template; and

(d) contacting the products of step (c) with RNA polymerase to generate an RNA library, said library being transcribed from said second nucleic acid strand.

2. The method of claim 1, wherein said method is used to introduce one or more mutations into said library.

3. A method for reducing sequence variation in a population of nucleic acid molecules, said method comprising:

(a) providing a first population of single-stranded nucleic acid templates of varying sequence, each of substantially all of said templates

comprising a coding sequence and an operably linked promoter sequence;

(b) hybridizing to the members of said first population a second population of substantially complementary single-stranded nucleic acid fragments, said fragments being shorter in length than said nucleic acid template and said fragments being of substantially identical sequence;

(c) contacting the hybridization products of step (b) with both a DNA polymerase which lacks strand displacement activity and a DNA ligase under conditions in which said fragments act as primers for the completion of a second nucleic acid strand which is substantially complementary to said nucleic acid template; and

(d) contacting the products of step (c) with RNA polymerase to generate a population of RNA molecules, said population of RNA molecules being transcribed from said second nucleic acid strand and having reduced sequence variation relative to said first population of single-stranded nucleic acid templates.

4. The method of claim 3, wherein said method is used to remove one or more mutations from said first population of single-stranded nucleic acid templates.

5. The method of claim 3, wherein step (b) comprises hybridization of said first population of single-stranded nucleic acid templates to two or more different populations of substantially complementary single-stranded nucleic acid fragments.

6. The method of claim 1 or 3, wherein said mixture of substantially complementary single-stranded nucleic acid fragments is generated by cleaving

a double-stranded nucleic acid molecule or by synthesis of random oligonucleotides.

7. The method of claim 1 or 3, wherein said single-stranded nucleic acid template is generated using an M13 phage carrying said nucleic acid, by digestion of one strand of a double-stranded nucleic acid template using gene VI exonuclease or lambda exonuclease, by capture of a biotinylated single nucleic acid strand using streptavidin, or by reverse transcription of RNA.

8. The method of claim 1 or 3, wherein said mixture of substantially complementary single-stranded nucleic acid fragments comprises at least about 100 different species of nucleic acid fragments.

9. The method of claim 1 or 3, wherein step (b) is carried out using between 1 and approximately 1000 fragments per single-stranded nucleic acid template.

10. The method of claim 1 or 3, wherein a single strand of the product of step (c) is used as a nucleic acid template and steps (b) and (c) are repeated.

11. The method of claim 10, wherein said steps (b) and (c) are repeated, using, in each round, the product of step (c) as said nucleic acid template.

12. The method of claim 1, wherein said method further comprises providing one or more single-stranded nucleic acid fragments which form a

homoduplex with said single-stranded nucleic acid template and carrying out step (b) in the presence of said homoduplex-forming fragments.

13. The method of claim 1 or 3, wherein said promoter is a T7 promoter.

14. The method of claim 1 or 3, wherein said DNA polymerase is T4 DNA polymerase.

15. The method of claim 1 or 3, wherein said method further comprises amplifying said product of step (c) prior to said contacting step (d).

16. The method of claim 1 or 3, wherein said method further comprises the step of:

(e) translating said RNA library to generate a protein library.

17. The method of claim 1 or 3, wherein said method further comprises the step of:

(e) linking to the 3' terminus of said coding sequence of each of substantially all of the members of said RNA library an amino acid acceptor molecule.

18. The method of claim 17, wherein said method further comprises the step of:

(f) translating said RNA library to generate an RNA-protein fusion library.

19. A method for generating a nucleic acid library, said method comprising:

(a) providing a population of single-stranded nucleic acid templates, each of said templates comprising a coding sequence;

(b) providing a population of single-stranded nucleic acid molecules of varying sequence, said population of single-stranded nucleic acid templates and said population of single-stranded nucleic acid molecules of varying sequence being substantially complementary;

(c) hybridizing said population of single-stranded nucleic acid templates with said population of single-stranded nucleic acid molecules of varying sequence under conditions sufficient to form duplexes; and

(d) contacting said duplexes with one or more excision/repair enzymes under conditions that allow said enzymes to correct mismatched base pairs in said duplexes.

20. The method of claim 19, wherein said method further comprises providing a population of single-stranded templates derived from the product of step (d) and repeating steps (c) and (d).

21. The method of claim 20, wherein said steps (c) and (d) are repeated, using, in each round, a population of single-stranded templates derived from the product of step (d).

22. A method for generating a nucleic acid library, said method comprising:

(a) providing a population of single-stranded nucleic acid templates, each of said templates comprising a coding sequence;

(b) hybridizing to said population of single-stranded nucleic acid templates a mixture of substantially complementary single-stranded nucleic acid fragments, said fragments being shorter in length than said nucleic acid template;

(c) contacting each of the hybridization products of step (b) with both a DNA polymerase which lacks strand displacement activity and a DNA ligase under conditions in which said fragments act as primers for the completion of a second nucleic acid strand which is substantially complementary to said nucleic acid template; and

(d) contacting the products of step (c) with one or more excision/repair enzymes under conditions that allow said enzymes to correct mismatched base pairs in said products.

23. The method of claim 22, wherein said method further comprises providing a population of single-stranded templates derived from the product of step (d) and repeating steps (b) - (d).

24. The method of claim 23, wherein said steps (b) - (d) are repeated, using, in each round, a population of single-stranded templates derived from the product of step (d).

25. The method of claim 19 or 22, wherein said contacting with said excision/repair enzymes is carried out in vivo.

26. The method of claim 25, wherein said contacting with said excision/repair enzymes is carried out in a bacterial cell.

27. The method of claim 19 or 22, wherein said contacting with said excision/repair enzymes is carried out in vitro.

28. The method of claim 19 or 22, wherein said single-stranded nucleic acid template is generated using an M13 phage carrying said nucleic acid, by digestion of one strand of a double-stranded nucleic acid template using gene VI exonuclease or lambda exonuclease, by capture of a biotinylated single nucleic acid strand using streptavidin, or by reverse transcription of RNA.

29. The method of claim 19 or 22, wherein step (b) is carried out using between 1 and approximately 1000 single-stranded nucleic acid molecules of varying sequence or single-stranded nucleic acid fragments per single-stranded nucleic acid template.

30. The method of claim 19 or 22, wherein said method further comprises the step of:

(e) amplifying said product of step (d).

31. The method of claim 19 or 22, wherein each of said coding sequences is operably linked to a promoter sequence.

32. The method of claim 31, wherein said method further comprises the step of:

(e) transcribing the products of step (d) to generate an RNA library.

33. The method of claim 32, wherein said method further comprises

the step of:

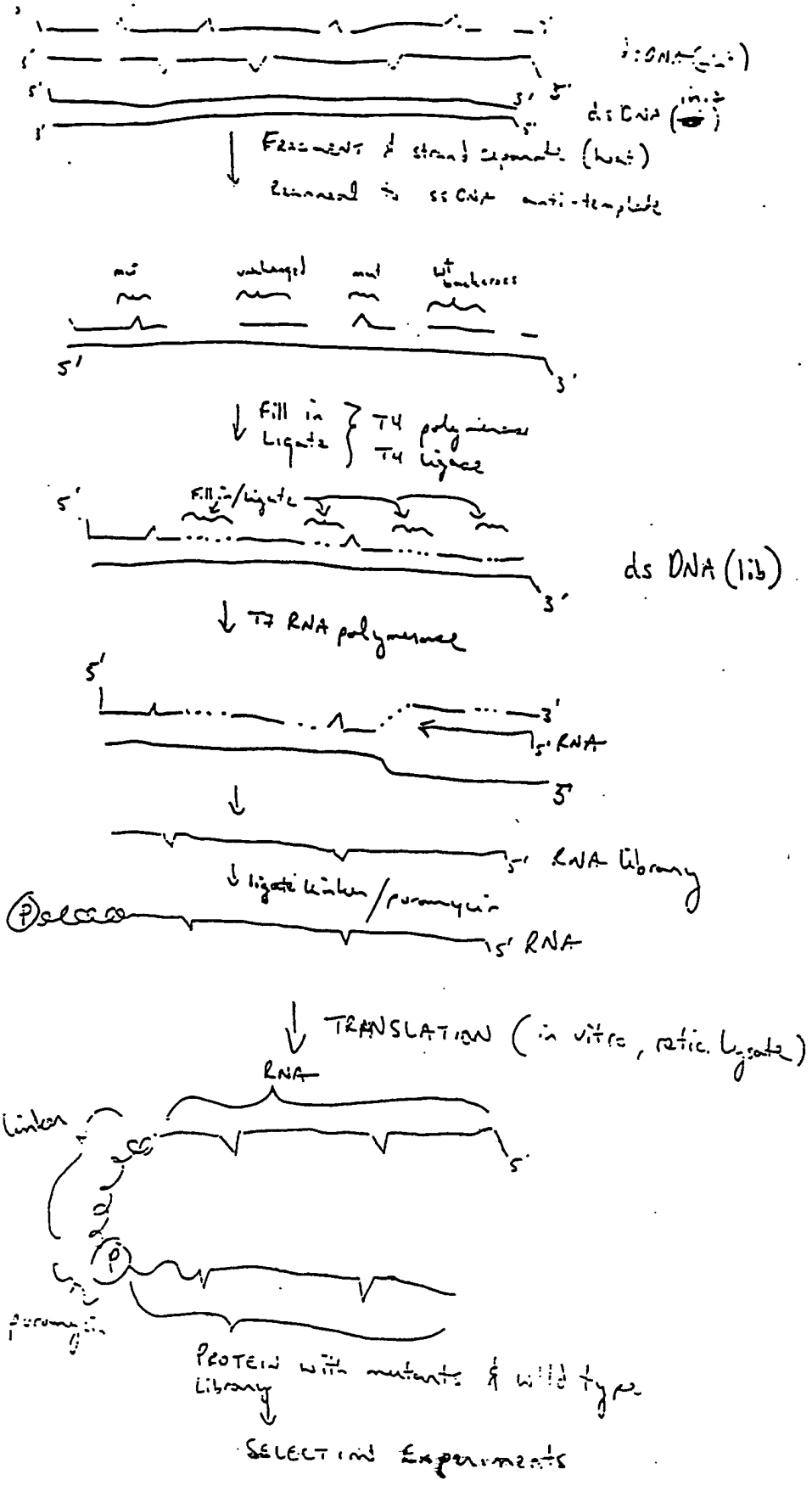
(f) translating said RNA library to generate a protein library.

34. The method of claim 32, wherein said method further comprises the step of:

(f) linking to the 3' terminus of said coding sequence of each of substantially all of the members of said RNA library an amino acid acceptor molecule.

35. The method of claim 34, wherein said method further comprises the step of:

(g) translating said RNA library to generate an RNA-protein fusion library.


ds DNA (in vitro)
Fragment & strand separate (heat)
Renatured to ss DNA anti-template
mut
wt
mut
wt
5'
3'
Fill in
Ligate
T4 polymerase
T4 ligase
Fill in/ligate
ds DNA (lib)
T7 RNA polymerase
RNA
RNA library
ligate linker/puromycin
TRANSLATION (in vitro, retic. lysate)
linker
puromycin
PROTEIN with mutants & wild type
Library
SELECTION Experiments

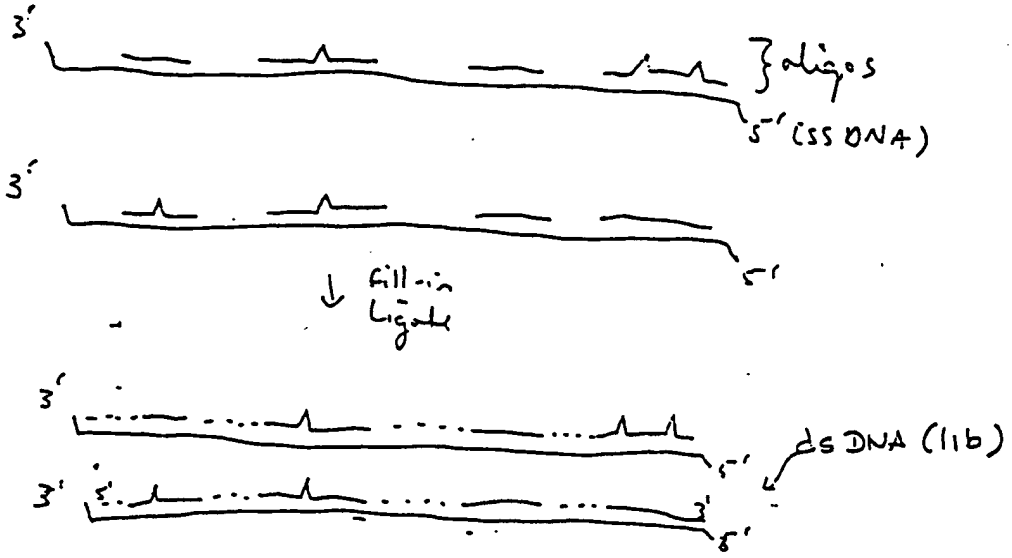


FIGURE 2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US99/14776

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :C12P 21/06, 19/34; C12N 15/74

US CL :435/69.1, 91.1, 91.2, 91.21, 472

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/69.1, 91.1, 91.2, 91.21, 472

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

NONE

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

STN, WEST

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	ALBERTS et al. Recombinant DNA Technology. In: Molecular Biology of Cell. New York: Garland Publishing, INC., 1994, pages 291-312, entire document..	1-35
Y	MATTHEWS et al. Biochemistry. Redwood City: The Benjamin/Cummings Publishing Company, 1990, pages 826, 827, 122-127 and 902-909 , entire document.	1-35
Y	US 5,580,730 A (OKAMOTO) 03 December 1996, abstract.	1-35
Y	US 4,994,379 A (CHANG) 19 February 1991, abstract.	1-35

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
17 SEPTEMBER 1999	22 OCT 1999

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US99/14776

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 5,447,839 A (MANOS et al) 05 September 1995, abstract.	1-35
Y	KACZOROWSKI et al. C0-operativity of hexamer ligationl. Gene. 1996, Vol. 179, No. 1, pages 189-193, especially the abstract.	1-35
Y,P	MINTZ et al. EHDI-An IH-domain-containing protein with a specific expression pattern. Genomics. 01 July 1999, Vol. 59, No. 1, pages 66-71.	1-35